Terabits Networks for Extreme Scale Science Workshop

Terabits Backbone Networking Challenges Session

Session Overview and Mission

February 16-17, 2011
Rockville, Maryland

DOE

Office of Science

Welcome and Introduction

- Session Facilitators
  - Inder Monga (ESnet)
  - Tom Lehman (USC/ISI)
- Session Scribe
  - Volunteers needed!
  - https://docs.google.com/document/d/1gE_RHOswqiqmowZavtp-pZSbpGhZVEuPxbJP_CQ17xQ/edit?hl=en&authkey=CNbEzpIC
- Workshop Web Site
  - https://indico.bnl.gov/internalPage.py?pageId=0&confId=319

Welcome and Introduction

- Four 1.5 hour Breakout Sessions over two days
- Input:
  - 5 planned presentations
  - 5 Questions distributed over 6 hours of discussion time
    - Networking fundamentals at extreme scale
    - Multi-domain, Multi-layer control and provisioning
    - Federated network services and performance monitoring
    - Network Testbeds and experimental infrastructure
    - Building a consistent and seamless end-to-end view
  - Open discussion (but on topic)

Welcome and Introduction

- Outputs
  - Identification of the key issues, challenges, and research issues that DOE should focus on
    - In anticipation of Terabits scale networks in the next 4-8 years
    - Within the context of the DOE mission and associated user requirements (end-to-end networking to enable Extreme-scale applications)
  - Workshop report with sufficient level of description to be used as a reference for DOE future activity planning
    - May need your help after the workshop to write and/or review portions of the report

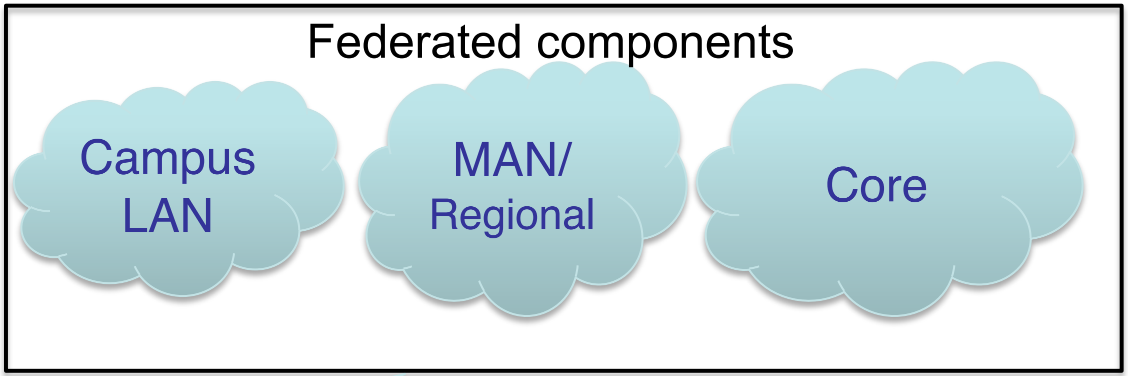
Discussion Structure and Objectives

- Questions are posed, please stay on topic
  - Chairs may interrupt when not on topic
- Open and free discussion is the objective
- Bold ideas and opinions welcome
- Ideas and thoughts do not need to be completely thought out or 100% defined
- Discuss, evaluate, provide opinions, and advocate for vision of the future as it relates to each of the topic areas
- Not trying to achieve a group consensus, but looking for key problems or future directions
  - Extract key research topics from the various future possibilities

Session Goals

Visualization

Raw Data Sources



Federated components

Campus LAN

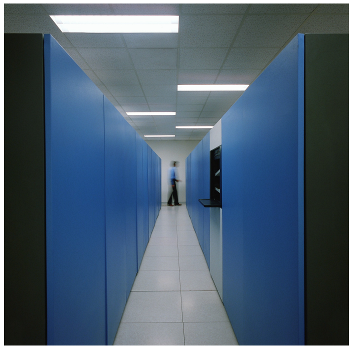
MAN/ Regional

Core

Data Stores

Enabling the Terabit network infrastructure within DOE app. context:

- **End-to-end focus**
- **Multi-layer**



Exascale Computation

Session Goals

- Identify the challenges to the development and deployment of a Terabit Networking infrastructure in the DOE context within 4-8 years. The information presented in the opening general session and from the reading materials should be used as guidance.
- Terabit networking should be considered in the context of end-to-end networking technologies with consideration of End Systems as well as Core, Regional, Campus, and Site networks.
- Summarize the impact of the future User/ Application requirements (including Exascale Computing and Zetabyte Data) on future networks.

Session Goals (cont'd)

- Describe possible technologies and architectures to realize this end-to-end capability.

- Identify the key research and development areas

- General questions applicable to all topic areas:
  - What are the possible futures we might see as things are evolving now?
  - What are the features in current networks we want to keep?
  - What are the features missing in current networks we want to add?
  - What are the possible futures we should try and create given the User/Application requirements?

Schedule

- Breakout Session 1: 13:20 – 14:50
  - Recent Advances in Network Systems - **Drew Perkins**
  - Question 1: Revisiting network fundamentals at extreme scale
  - Question 2: Multi-layer, multi-domain network provisioning
- Breakout Session 2: 15:05 – 16:35
  - Recent Advances in Flow Switching and Control: OpenFlow – **Rob Sherwood**
  - Question 2: Multi-layer, multi-domain network provisioning
  - Question 3: Federated Network Services, Management, and Performance Monitoring

Schedule

- Breakout Session 3: 09:00 – 10:30
  - End-to-end Network Performance Monitoring – **Jason Zurawski**
  - Question 3: Federated Network Services, Management, and Performance Monitoring
  - Question 4: Testbeds and Experimental Network Infrastructures
- Breakout Session 2: 10:45 – 12:15
  - ESnet Table Top Testbed – **Brian Tierney**
  - GENI Project – **Aaron Falk**
  - Question 4: Testbeds and Experimental Network Infrastructures
  - Question 5: Extension of Core Network Capabilities to End Systems for Massive Data Movement

Question 1 – Revisiting network fundamentals at extreme scale

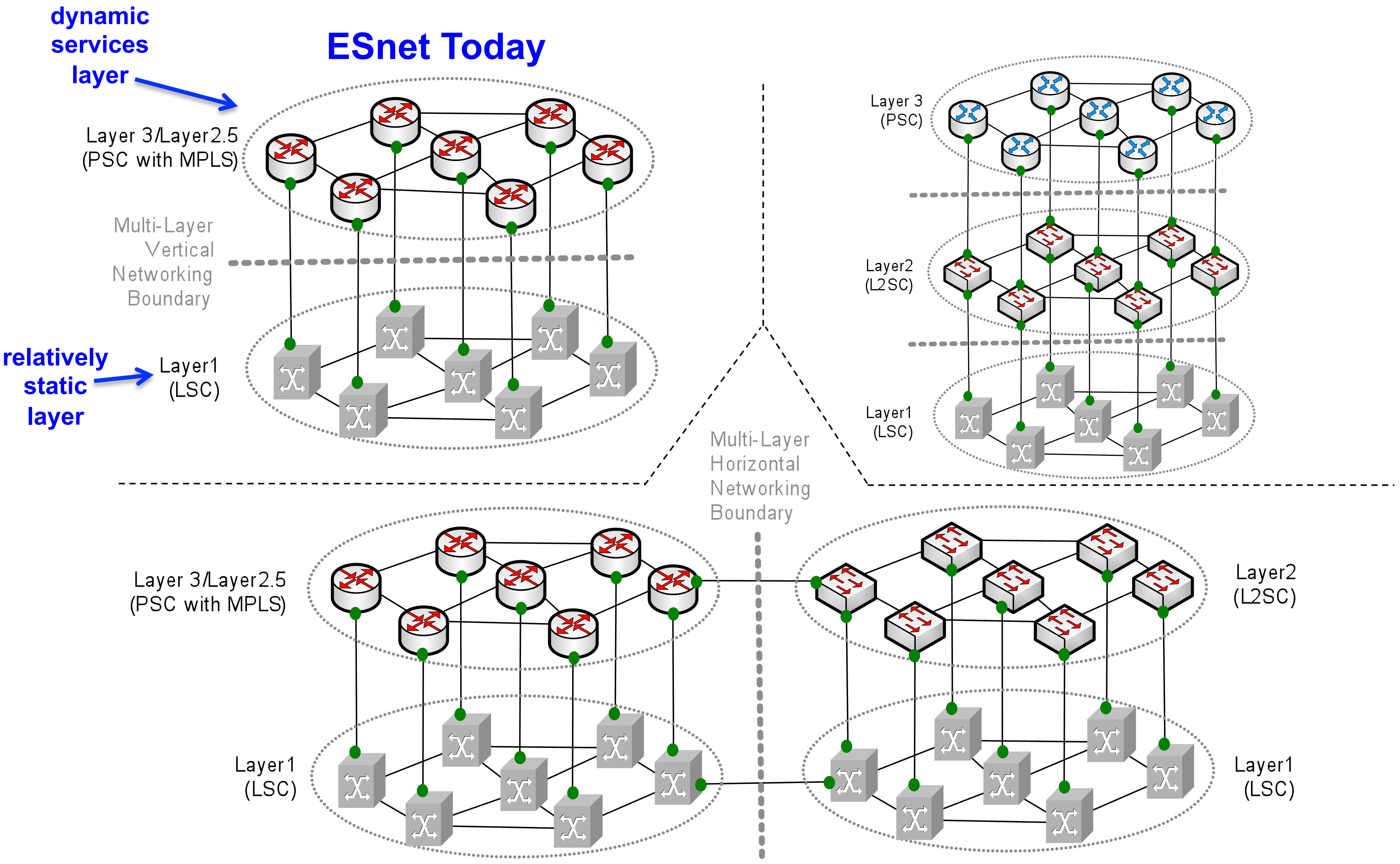
- Fundamental technical challenges to scale the network 1000x and end-systems by 100x
  - Network Systems, Architecture and Protocols
  - Packet-switching, Circuit Switching or Flow switching
  - Traffic Engineering and Dynamic provisioning
  - Network management, control and provisioning - in an end-to-end federated environment (Campus to Core)
  - Impedance mismatches and mid-boxes between network domains including end-systems

Question 1 – Revisiting network fundamentals at extreme scale

- What are the fundamental technical challenges of scaling existing networks by 1000x in the core and 100x in the end systems to realize federated end-to-end terabits networking in the context of the following backbone networking concepts:
  - a) Network architecture and protocols
  - b) Packet-switching and circuit-switching
  - c) Network provisioning and traffic engineering
  - d) Federated network operation, control, and management
  - e) Impedance mismatch between core network and end systems effective bandwidth

Many Multi-Layer Topology Options

What will future DOE networks look like?



ESnet Today

dynamic services layer

Layer 3/Layer2.5 (PSC with MPLS)

Multi-Layer Vertical Networking Boundary

relatively static layer

Layer1 (LSC)

Layer 3 (PSC)

Layer2 (L2SC)

Layer1 (LSC)

Multi-Layer Horizontal Networking Boundary

Layer 3/Layer2.5 (PSC with MPLS)

Layer1 (LSC)

Layer2 (L2SC)

Layer1 (LSC)

● technology adaptation points

Question 1 – Discussion Topic Focus Summary

- Current DOE network technology/architecture is IP routers over WDM
- What are recommendations for DOE network technology/architecture when we have terabit networks?
  - should networks be multi-layer, with highly dynamic control at all the layers?
  - should networks be collapsed to a single technology layer
  - do networks need to be "virtualized" like clouds are doing with hosts?
  - what about boundary between wide area and regional/site networks? are there connections/peering at multiple layers?
  - What is the right architecture at campuses to handle multi-100G traffic and aggregate/feed it to the terabit core? where are bottlenecks?
- What are the technology/architecture options for building terabit capacity networks. What are the key technology issues, challenges, and unknowns?
  - serialized terabit links? 10x100Gbps links?
  - any disruptive technologies in this timeframe?

Question 2: Multi-layer, multi-domain network provisioning

- The DOE operates a complex science environment consisting of science resources and national and international collaborators the generated distributed high-end science applications, each with unique network and cyber security requirements. What are the technical challenges of engineering multi-domain federated terabit network with differentiated capabilities to address the unique network requirements of different science applications in the following contexts?

Question 2: Multi-layer, multi-domain network provisioning (cont'd)

a) Multi-modal networking (packet-switching, dynamic, circuits, flow switching, etc.,) over a common federated terabits networks

b) Multi-domain cross-layer optical network virtualization

c) Inter-domain policy coordination and SLA for controlling and managing end-to-end dynamic circuits, virtual networks, network monitoring, cyber security, etc.

d) others

Question 2 - Discussion Topic Focus Summary

- Multi-layer internet architecture is necessary but has led to “competition” of intelligence between layers
- Multi-domain is the de-facto scenario going from campus - core, national scale - international partners
- Is multi-layer not optimized, and should all intelligence be focused on a single layer?
  - IP, Optical, Ethernet??
- What is the switching decision-making granularity dominance for terabit scale – flow based or packet based?
- What are the control mechanisms of the future? (G)MPLS, Openflow, OSCARS, point-and-click provisioning, active programmable networks?
- What are the right federation layer for multi-domain - at IP, at service layer, at MPLS-TP/Ethernet layer?

Question 3: Federated network services, management, and performance monitoring

- DOE network infrastructures is a complex collection of autonomous local, national, and international network systems that collaborate to deliver end-to-end performance several order of magnitude higher than what commercial best effort IP internet offer What the challenges of development new network management and performance tools and service or scaling existing ones to take work efficiently in end-to-end federated hybrid terabits networks with cross-layer support. Critical factor for consideration:

Question 3: Federated network services, management, and performance monitoring (cont'd)

- a) Multi-layer capabilities accessible by scientists at the application level and by network engineers the network layer 1-3
- b) Address inter-domain, access control, flow manipulation and control, policy coordination, etc., issues
- c) Predictive performance, real-time, post-Mortem, and x automated capabilities
- a) Other new control and management technologies

Question 4: Testbeds and experimental network infrastructure

- Given that terabits networks will require new and sometime radically different networking architecture, protocols, hardware, software, etc., that could be disruptive to existing production networks; What the role and benefits of testbeds and experimental networking at scale in the development of end-to-end terabit networking? Issues of considerations:
  - a) End-to-end testbeds addressing multi-layer, application prototyping, control plane, performance monitoring ,and inter-domains issues

Question 4: Testbeds and experimental network infrastructure (cont'd)

- b) Coordination of multi-agency testbeds
- c) Simulation-based network experimentation that leverage petascale and exascale computing
- d) Accelerating advances in extreme-scale networking through industry and vendor partnerships in testbed and experimental partnership.

Question 5: Extension of core network capabilities to end systems for massive data movement

- moving massive data generated by extreme-scale computing and large scientific instrument is a critical priority for DOE network infrastructures. What are the challenges of extending the abundant optical bandwidth made possible in the core by the emerging 100 GigE technologies to support the distribution of massive science data over very long distances? Potential issues for consideration:

Question 5: Extension of core network capabilities to end systems for massive data movement (cont'd)

- b) Transport protocols that can deliver 100x performance to high-end science application
- c) Harnessing parallel file systems features, multi-core host architectures, and I/O concurrencies to improve end-to-end data transfer throughputs
- d) Co-scheduling of network, storage, applications, etc., to maximize throughput and efficiency of data transfer

Other Questions we should consider?

Backup Material

Detailed Topics for Each Question Area

Question 1 - Detailed Topic Areas

- Path to terabit capacity
  - If optical line transmission is not the problem, where is it?
  - Client interfaces - path beyond 100GE? Challenges?
  - Issues of Aggregation? Forwarding? Buffering? LAG?
- Network Equipment and Architecture
  - Routers, Optical transport, switches or converged devices?
  - Scalable, cost-effective and energy-efficient system - is this optimizable?
- Flows, packets or circuits - what is the switching and add/drop quanta? How does it change from end-to-end?
- Optical Packet switching - how far are those solutions?
- Silicon Photonics - path to ubiquitous solutions?

Question 1 - Detailed Topic Areas

- What do the network elements of the future look like?
  - Link Speeds?
  - Chassis Capacity?
  - Flow Scaling?
  - What technologies dominate?
  - Cost?
  - Energy Usage Considerations?
- What will be the “best” approach to get terabit networking capacity in the next five years between wide area network sites – 10x100 or 1 x 1TB or 3x400G?

Question 1 - Detailed Topic Areas

- Where is the biggest challenge to get 1TB WAN connectivity? Can packet processing keep up – limited to what capacity or new architectures?
- What are the new hardware technologies in the Campus and Regional networks that will influence the utilization of the WAN or device access for Network services?
- Where are the limitations in the networking platform to accomplish high packet processing – compute chipsets, ASICS, memory, CAMs, surface area on boards, switch fabric?

Question 1 - Detailed Topic Areas

- Power and size are becoming important limitations when designing a next-generation device – how do future network innovations affect this? Do they make this problem worse? What innovations / research is needed to have higher capacity devices within a reasonable power/space budget?
  - What is the expectation for scaling?
  - Router with 100 1 TB links or just one 1TB uplink
- Is it time to focus again in Optical Burst and packet switching or we should leave that in the dream world still?

Question 1 - Detailed Topic Areas

- Do we see integration of all discrete components – electrical, optical processing under one physical chassis with cross-layer capability? What are the limitations of such an integrated device?
- What role does the new generation of ROADMS with filterless, colorless, directionless capabilities and WSS switches play in enabling on-demand networking? How do they affect future network architectures?
- "Packet switching at the edges, circuit switched all around" – is that statement going to be true in the next 5 years? Where is that Edge going to be – Host, LAN, MAN border?

Question 1 - Detailed Topic Areas

- What is the dominant data movement mechanism in the core – packet or circuit switched?
- What would be the major change for network technologies in the next five years?
  - What are some disruptive technologies right on the horizon
- Is call for energy-efficiency effectively change the design of networking hardware? What are the expected changes?
- What are the research areas or broad topics that may revolutionize the evolution of the network hardware?
  - Silicon Optics, Parallel multicore OS's, ….

Question 2 - Detailed Topic Areas

- Define the future architectures and requirements
  - End-Device(compute/storage/viz), Campus, Regional, WAN
- What is the switching decision-making granularity dominance – flow based or packet based? Will flow-based mechanisms favor circuit switched approaches more than packet-approaches?
- What are the impedance mismatches between Host and Campus, Campus and Metro, and Metro and WAN? Bandwidth, control, performance, visibility, troubleshooting. What needs to be researched/fixed?

Question 2 - Detailed Topic Areas

- What network architecture elements are necessary to assure guaranteed end-to-end performance?
- What are the preferable architectures of the future where convergence is headed? Multi-layer, single-layer, single-chassis?
  - All IP, All Optical, All Ethernet??
- What are the control mechanisms of the future? (G)MPLS, Openflow, OSCARS, point-and-click provisioning, active programmable networks?
- What architectures make it easier for “customers” of the network to use them?

Question 2 - Detailed Topic Areas

- What exactly is meant by "clean slate" design? and is it possible?
- How do you provide guaranteed packet delivery with a protection/restoration mechanism for terabit networks?
- What different technologies and architectures will differentiate LAN/WAN/MAN? Is there going to be a boundary or does it all end up at the host? Is the control plane going to be different or the same?

Question 2 - Detailed Topic Areas

- What are the advantages of divorcing the control plane and the hardware-forwarding plane? What are the pros and cons of such network architecture? Is there a larger role for open-source with this model?

Question 2 - Detailed Topic Areas

- What are timescales of network changes required in the future architectures? Where does the dynamic networking end – at the TCP layer, IP layer, Virtual circuit layer, WDM layer or needs to exist at all layers? What is the impact of dynamic lower layers on network architecture?
- How is network protection and recovery handled at Terabit/s
- Does a new distributed architecture needs to be investigated?

Question 2 - Detailed Topic Areas

- Is it possible to identify potential network architectures of the future?
    - All optical + Openflow
    - All packet routing
    - Hybrid architecture - packet routing over optical
    - Hybrid architecture - mix of packet routing and circuit switching over optical or carrier ethernet
    - others?

Question 3 - Detailed Topic Areas

- What "network capabilities/functions" do the networks of the future provide?
- There is a vision for a network service plane that the user interacts with to get network services. All co-scheduling like middleware for compute/cloud services uses this network service plane for getting access to network resources.
  - How many people believe this is the way to go?
  - What are the network services?
  - What are the appropriate mechanisms to bring applications and networks closer?

Question 3 - Detailed Topic Areas

- How can applications automatically notify the network of issues?
- How can network service provide intelligent services to the applications/users?
- How can the network services enable the applications to improve their own performance?
- How can networks become autonomic? Self-correcting/debugging? What sort of tools and embedded intelligence are needed?
- What role to provisioning systems like Openflow and OSCARS have in user facing network services and management?

Question 3 - Detailed Topic Areas

- What is the right federation architecture for stitching inter-domain services? For providing SLA-based service? Difference between intra-domain and inter-domain tools/protocols/architectures.
- Scenario “ A user wants network to find its own performance problems, troubleshoot and fix itself while routing its traffic over an alternate trouble-free path” – how far are we away from this vision? How can we get there? Is it what we need to shoot for?
- Will services embedded in the network (like cloud and/or data storage) be a major component of future network architectures? if so how will it impact network design?

Question 4 - Detailed Topic Areas

- Common experimental testbed
  - What’s needed?
  - How do you compete and showcase at the same time?
- What is the value to
  - Industry
  - Researchers
- Who are the users?
  - Researchers?
  - Grad students?
  - Network operators?
  - Solution labs of industry?

Question 4 - Detailed Topic Areas

- Are Testbeds required in order for DOE to meets is research objectives?
- How do we identify the key research areas that specifically can benefit from a testbed environment?
- Should the testbed facilities be small and self organized by research groups who need them?
- or should testbed facilities be run more as a larger infrastructure which can be adapted to multiple research projects?

Question 4 - Detailed Topic Areas

- What are the key research area that testbeds should focus on:
    - wide area transmission systems?
    - metropolitan area network switching and routing systems?
    - network architectures and control systems?
    - end-to-end data flow and performance optimization?
    - federated network systems?
    - networks as a service?
    - embedded network services (like cloud computing, or data storage)?

Multi-Layer Networking

An Architecture Framework

DOE

Office of Science

Multi-Layer Network Architecture

- This work is an attempt to capture ongoing discussions regarding multi-layer networks within the DOE community (and beyond)
- The desire is to develop a framework for discussion
  - not intended to define or require anything specific regarding implementations or designs
- Intended audience is researchers and network system implementers in DOE

Multi-Layer Network Architecture

- Network architectures and implementations are on the edge of a transformation from a single-layer to a multi-layer paradigm.
- This is in the context of "network service provision and control"
  - physical construction has always been multi-layer
- In many ways, this represents an evolution from IP QoS, to MPLS, to "GMPLS" (below routers) network control and provisioning.

Multi-Layer Network Architecture

- Additional benefits of the multi-layer paradigm include:
  - higher efficiency (create bandwidth where needed)
  - lower costs
    - bandwidth is cheaper at the lower layer (e.g. cut-through routing)
  - service flexibility and network agility
    - can provide more then just IP service
  - better service guarantees
    - use the transport technology that best meets the request constraints (e.g. SONET, WDM for low jitter, etc)

Multi-Layer Network Architecture

- As we address the complexities of integrated "network service provision and control " across multiple network layers:
  - a common set of terminology/constructs will facilitate discussions, designs, and implementations
- Another goal of this document to provide a framework, context, and definition of where more detailed research and development efforts are needed.

Data Plane Descriptions

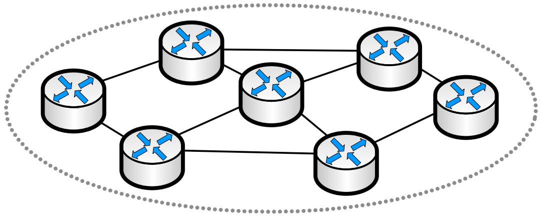
- For this architecture definition we identify the dataplane options in terms of “technology regions”.
- A “technology region” is a set of network elements which are grouped together and utilize the same dataplane "technology type".
- The technology types are defined using the standard Generalized Multi-Protocol Label Switching (GMPLS) nomenclature:
  - Packet Switching Capable (PSC) layer
  - Layer-2 Switching Capable (L2SC) layer
  - Time Division Multiplexing (TDM) layer
  - Lambda Switching Capable (LSC) layer
  - Fiber-Switch Capable (FSC)

Data Plane Descriptions

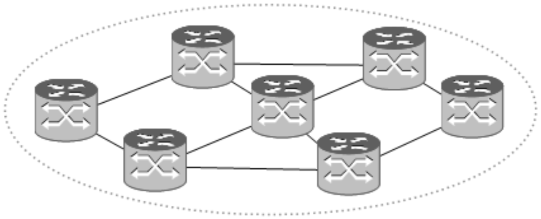
- Additionally, we can associate these technology types with the more common layer terminology:
  - Layer 3 for PSC (IP Routing)
  - Layer 2.5 for PSC (MPLS)
  - Layer 2 for L2SC (often Ethernet)
  - Layer 1.5 for TDM (often SONET/SDH)
  - Layer 1 for LSC (often WDM switch elements)
  - Layer 0 for FSC (often port switching devices based on optical or mechanical technologies)
- DataPlane - A set of network elements which receive, send, and switch the network data
  - For this architecture work we assume each technology region is of a single technology type. In reality a technology region may have multiple technology types.

Data Plane Technology Types/Regions

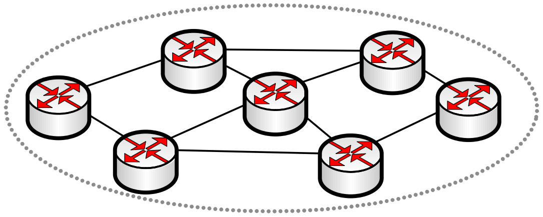
Layer 3
(PSC)



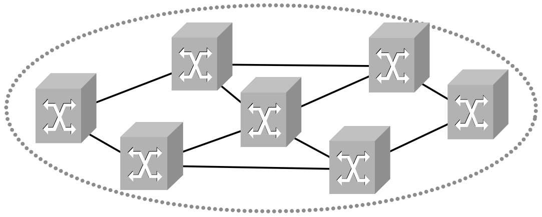
Layer1.5
(TDM)



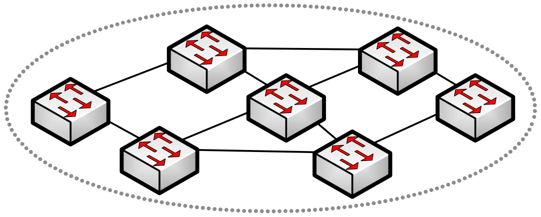
Layer 3/Layer2.5
(PSC with MPLS)



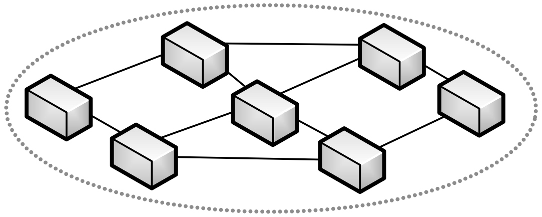
Layer1
(LSC)



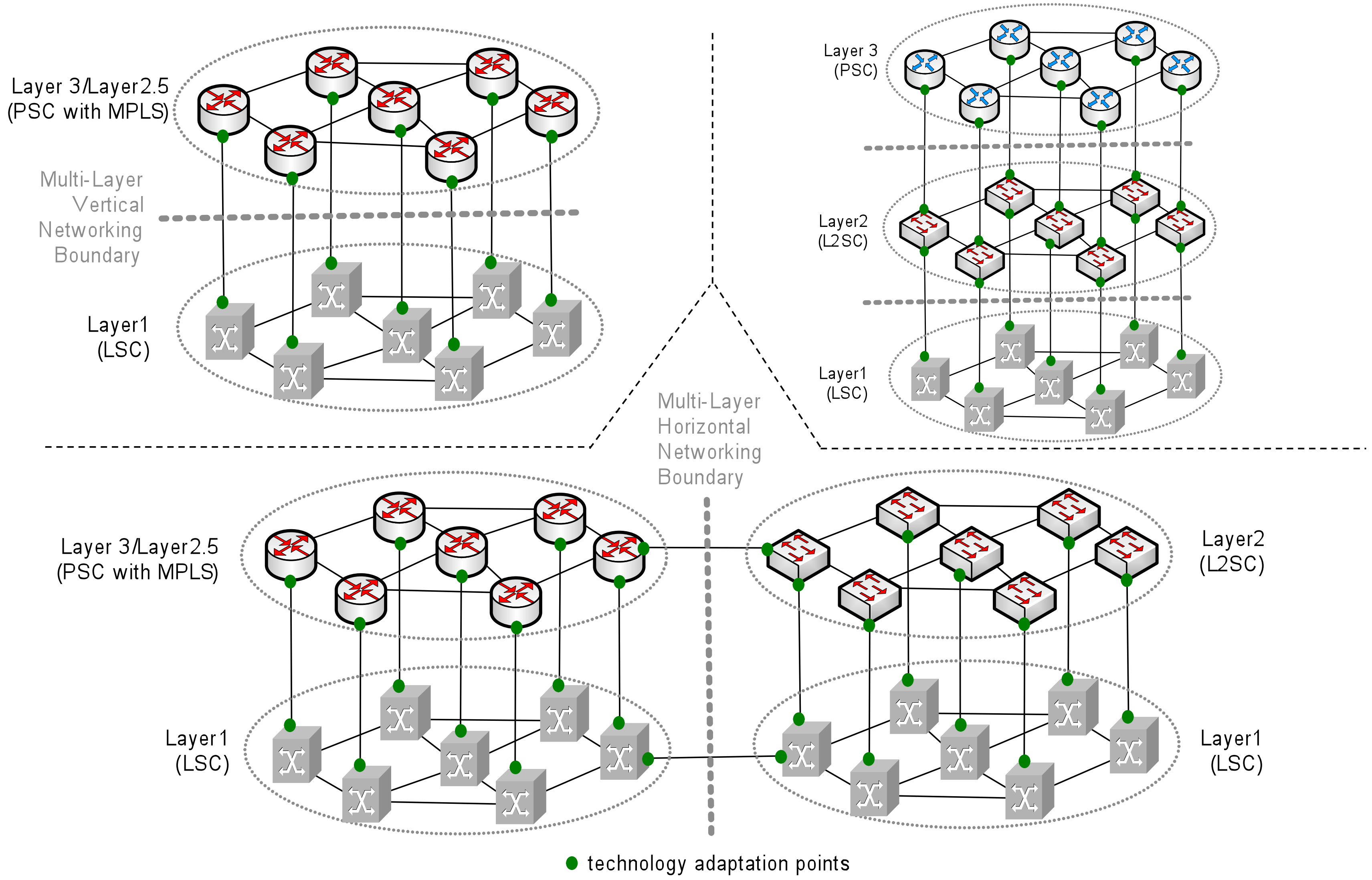
Layer2
(L2SC)



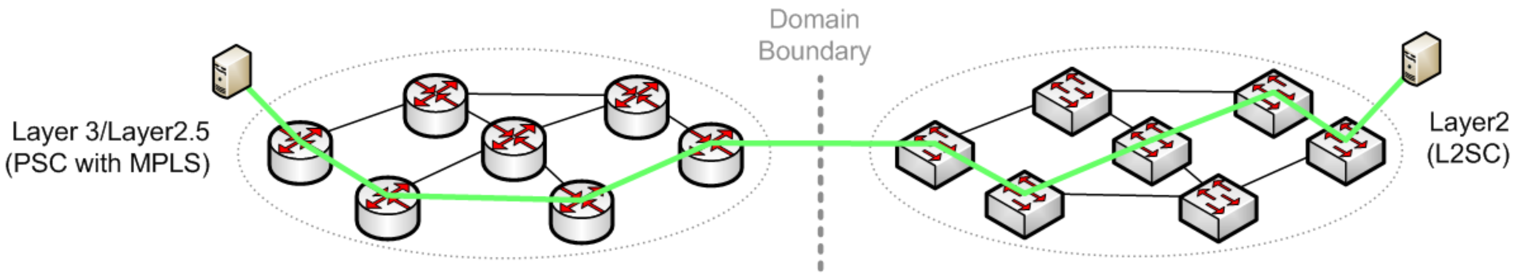
Layer0
(FSC)



Many Multi-Layer Topology Options

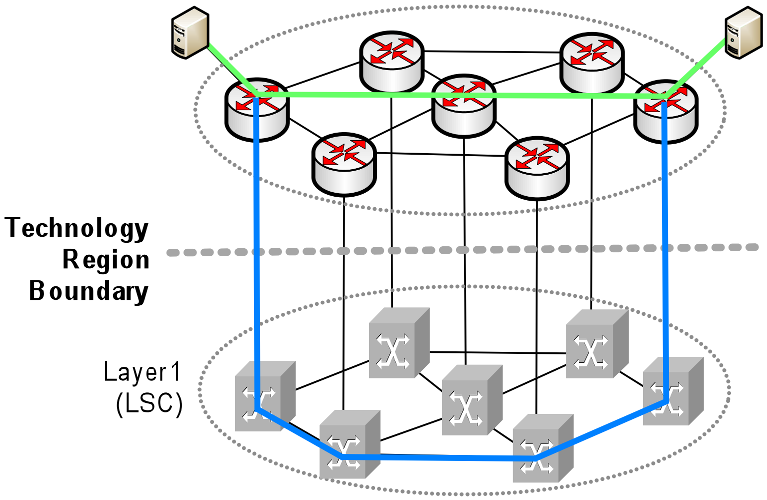


Current Network Provisioning



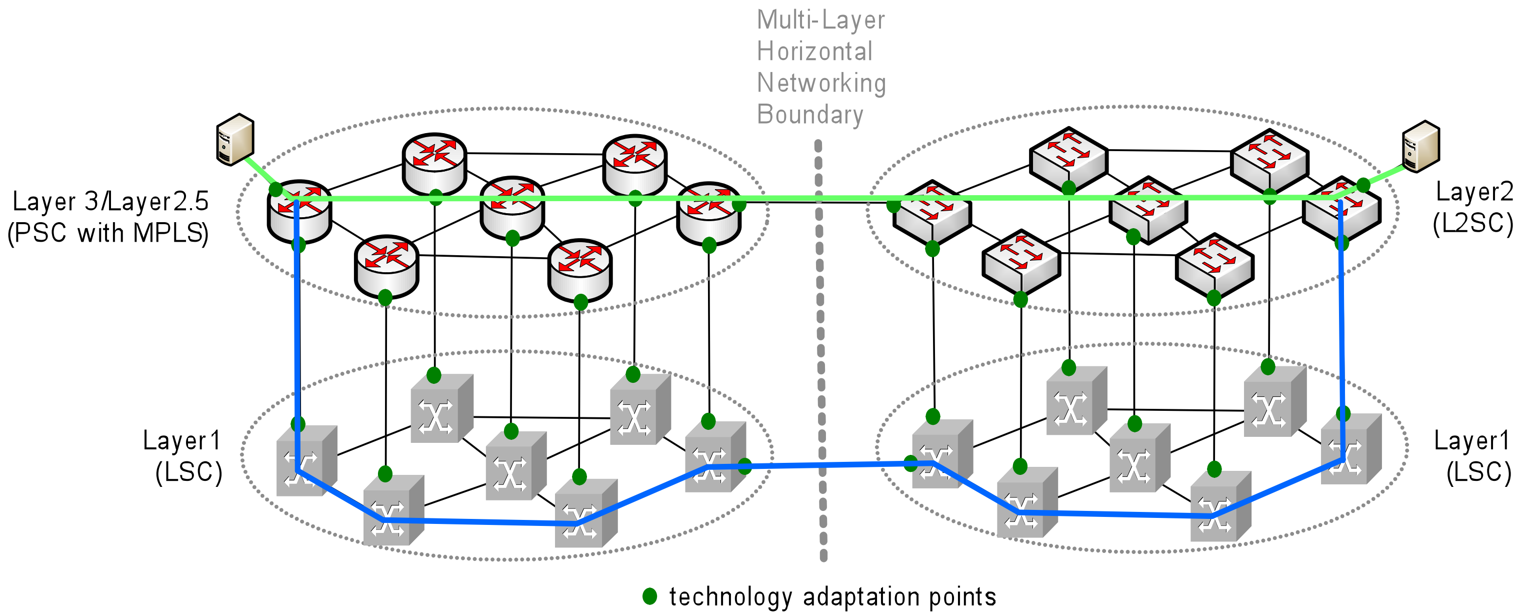
- Horizontal Inter-Domain Provisioning
- Single Layer Intra-Domain
- using systems like OSCARS, DRAGON

Vertical Intra-Domain is Desired



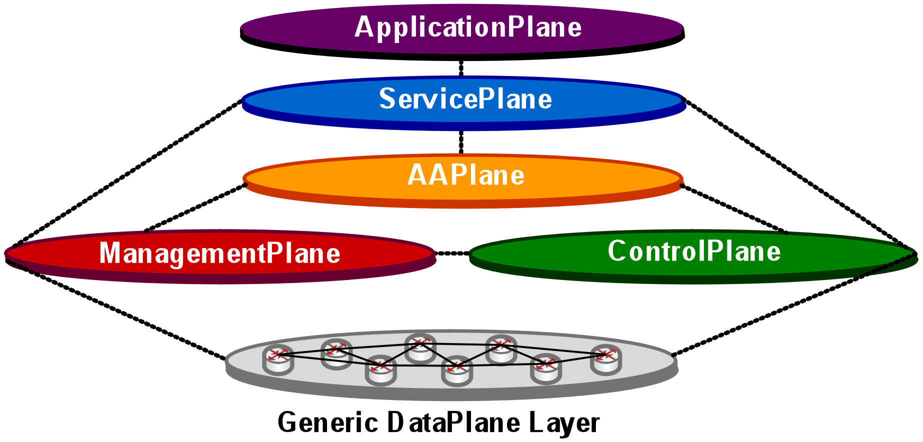
- Provision services at lower layer to create a topology element (link between routers) at the higher layer
- Subsequently provision any remaining bandwidth at the higher level
  - may provision a 10Gbps LSC link in response to a request for 5Gbps immediate need
  - remaining 5Gbps available for subsequent service requests

Vertical Inter-Domain Even Better

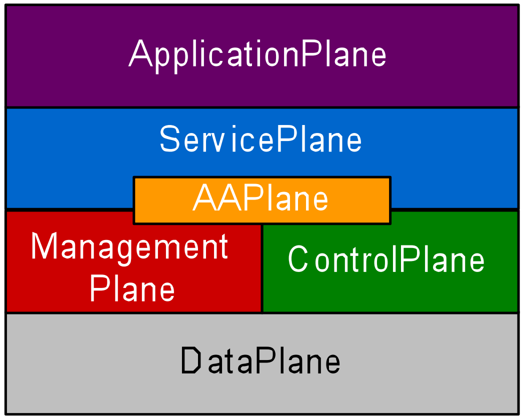


- Lower layer service provisioning may cross a domain boundary
- Networks may peer a multiple layers and flexibly provision services

Multi-Layer Architecture
Capability Planes

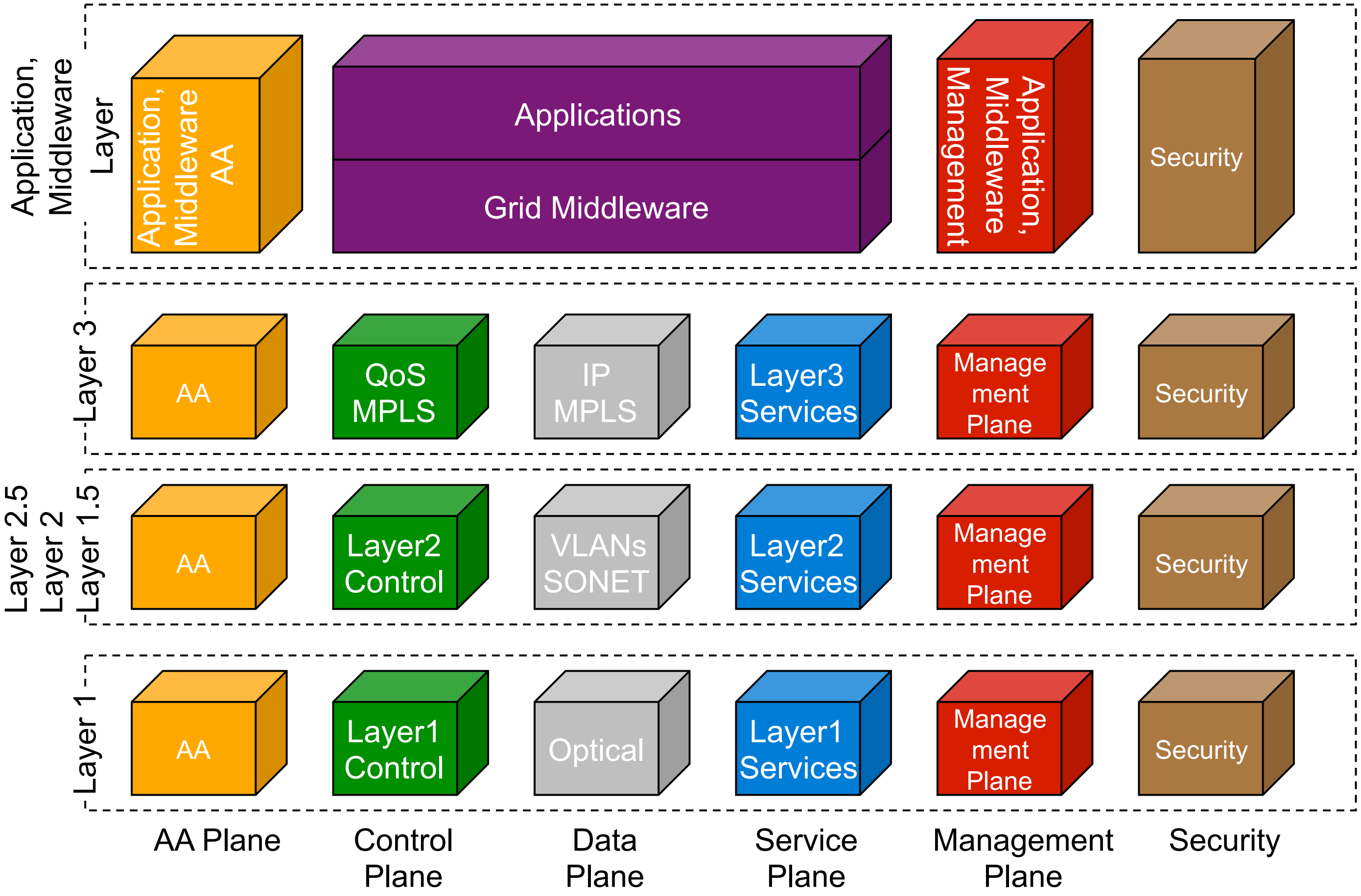


CapabilityPlanes - Graphical View



CapabilityPlanes - Functional Layered View

DataPlane Layer Centric View of Capabilities



	AA Plane	Control Plane	Data Plane	Service Plane	Management Plane	Security
Application, Middleware Layer	Application, Middleware AA	Applications Grid Middleware			Application, Middleware Management	Security
Layer 3	AA	QoS MPLS	IP MPLS	Layer3 Services	Manage ment Plane	Security
Layer 2.5 Layer 2 Layer 1.5	AA	Layer2 Control	VLANs SONET	Layer2 Services	Manage ment Plane	Security
Layer 1	AA	Layer1 Control	Optical	Layer1 Services	Manage ment Plane	Security

Multi-Layer Architecture
Capability Planes

- ***CapabilityPlanes*** are defined to capture the key areas required to manage and control a DataPlane layer.
- ControlPlane
  - routing, path computation, and signaling
  - maintaining topology information and configuring network elements in terms of data ingress, egress, and switching operations
  - one of two CapabilityPlanes which directly interacts with the DataPlane

Multi-Layer Architecture

Capability Planes

- ServicePlane
  - set of systems and processes that are geared towards providing services to users and maintaining state on those services
  - will generally rely on the functions of the ControlPlane and/or ManagementPlane to effect actual changes on the DataPlane
  - will typically maintain databases on current and future service instantiations and coordinate associated workflow processes

Multi-Layer Architecture
Capability Planes

- ManagementPlane
  - monitor, manage, and troubleshoot the network
  - includes functions to support user services as well as network maintenance, upgrades, and reconfigurations
  - responsible for collecting data and monitoring of the network
  - may also include capabilities to configure the network elements with the support of the control plane or via independent actions
  - one of two CapabilityPlanes which directly interacts with the DataPlane.

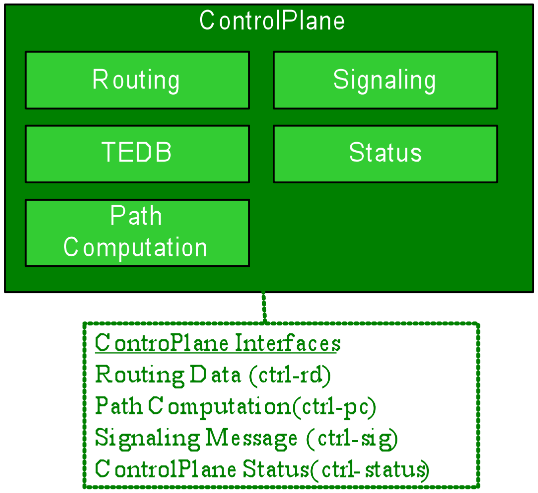
Multi-Layer Architecture

Capability Planes

- AAPlane (Authentication and Authorization)
  - responsible for the mechanisms which allow the other planes to identify and authenticate users and receive associated policy information
- ApplicationPlane
  - provides higher level functions which will generally be tailored for domain specific purposes.
  - relies on the capabilities offered by one or more ServicePlanes to accomplish its objectives
  - boundary between the ApplicationPlane and ServicePlane is the network demarcation point

Multi-Layer Architecture
Capability Planes

- For Each CapabilityPlane we define the following:
  - Functions
  - Function Interfaces
  - Layer Unique Considerations
  - Security Considerations
- For Example:



ControlPlane

Routing

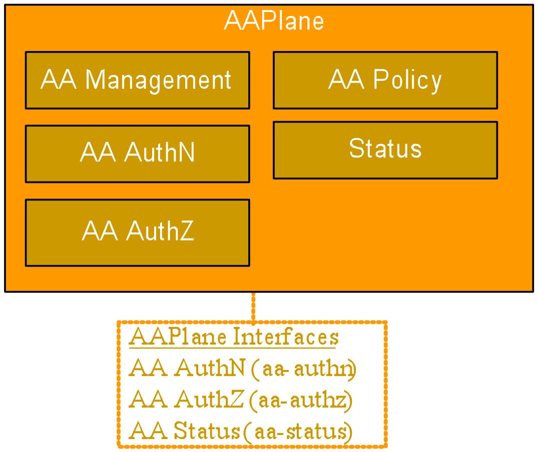
Signaling

TEDB

Status

Path Computation

ControPlane Interfaces
Routing Data (ctrl-rd)
Path Computation(ctrl-pc)
Signaling Message (ctrl-sig)
ControlPlane Status(ctrl- status)



AAPlane

AA Management

AA Policy

AA AuthN

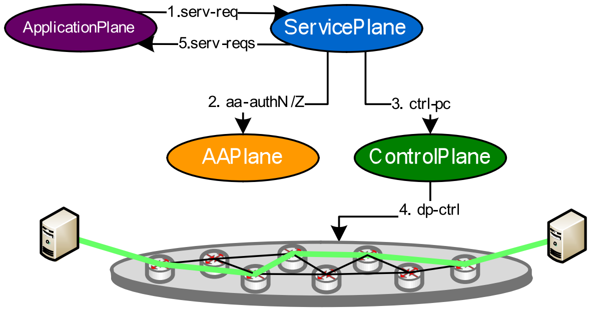
Status

AA AuthZ

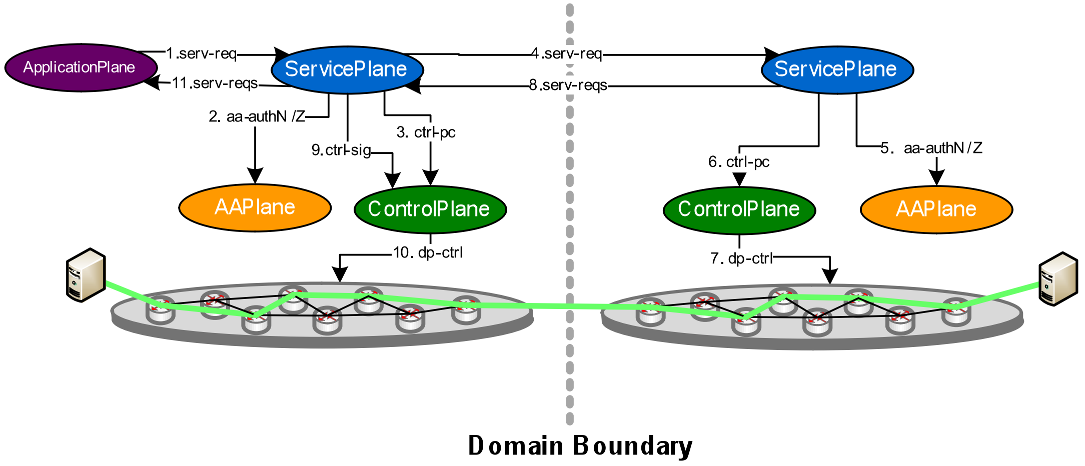
AAPlane Interfaces
AA AuthN (aa- authn)
AA AuthZ (aa-authz)
AA Status(aa-status)

WorkFlow Development

- Single Domain Circuit Request

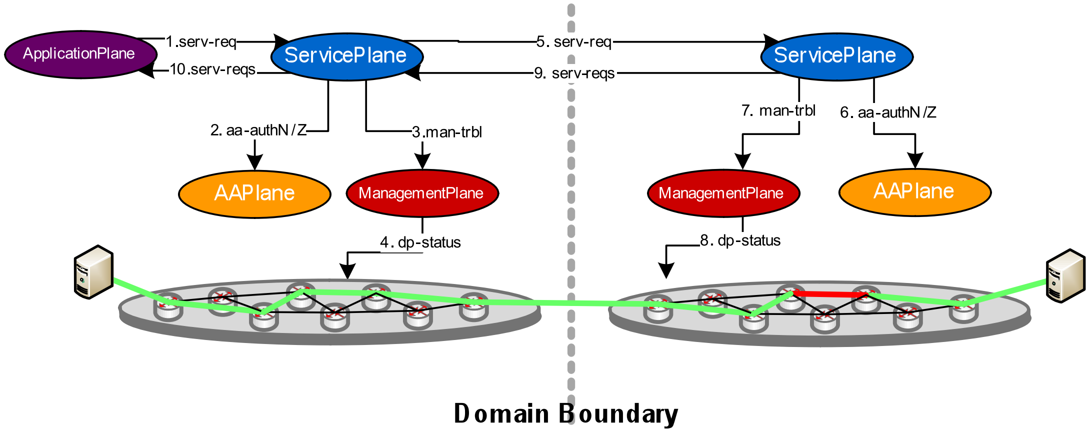


- Multi-Domain Circuit Request - InterDomain



WorkFlow Development

- Monitoring and Troubleshooting Workflow



- Many workflows possible including multi-layer integrated vertical provisiong scenarios

Multi-Layer Architectures and Service Concepts

- Multi-Layer Networking - Vertical
- Multi-Layer Networking – Horizontal
- Multi -Layer Networking – Combined
- Multi-Layer Networking – InterDomain
- Hybrid Networking (Multi-Layer Traffic Engineering and Traffic Grooming)
- Nested Capability Planes
- Discussed in architecture document
- http://hybrid.east.isi.edu/twiki/pub/HybridMLN/Pubs/multi-layer-architecture-v9.0.pdf